APPEL A PROPOSITIONS BRG – 2003 - 2004

TITRE DU PROJET
« ANALYSE ET PREDICTION DE PATRONS DE DESEQUILIBRE DE LIAISON DANS LES COLLECTIONS DE RESSOURCES GENETIQUES DE PLANTES ANNUELLES OU PERENNES, AUTOGAMES OU ALLOGAMES »

RAPPORT FINAL

DECEMBRE 2006

Responsable(s) scientifique(s) :

Marc Seguin

Cirad, Av. Agropolis, TA 80/03, 34098 MONTPELLIER, Cedex 5
Téléphone : 04 67 61 71 27, télécopie : 04 67 61 57 93, marc.seguin@cirad.fr

ASPECTS ADMINISTRATIFS ET OBJECTIFS DES RECHERCHES

ASPECTS ADMINISTRATIFS

Date d'engagement : 01/01/04

Montant du budget : 63 000 € + solde 2006 : 7 000 €

Cofinancements obtenus :

- ATP Cirad 22/2002 (blé, sorgho, hévéa) : 63 000 €, 2002-2005
- GABI Génoplante (sorgho) : nd, 2004-2005
- CGIAR, Challenge Program Generation (blé, riz, sorgho) : 100 000 €, 2005-2007
- ACI stagiaire post-doc, Ministère de la Recherche : 39 000 €, 18 mois, 2003 - 2005

Participants au projet :

Chercheurs permanents, coordinateurs plantes/équipes :

COURTOIS Brigitte [(1)] : riz,
DAVID Jacques [(2)] : blé dur
DEU Monique [(1)] : sorgho
RONFORT Joëlle [(2)] : *Medicago truncatula*
SEGUIN Marc [(1)] : hévéa

Chercheurs & techniciens permanents :

BATAILLON Thomas [(2)], BILLOT Claire [(1)], CHANTRET Nathalie [(2)], GLASZMANN Jean-Christophe [(1)], GLEMIN Sylvain [(3)], LE GUEN Vincent [(1)], TSITRONNE Anne [(2)], POMIES Virginie [(1)], WEBER Christelle [(1)]

Stagiaires (post-doc & master) :

ATTARD Agnès [(1)], CENCI Alberto [(2)], EL AZHARI Najoi [(1)], MAYNADIER Marie [(2)], HAUDRY Annabelle [(2)]

(1) UMR 1096, Polymorphismes d'Intérêt Agronomique, Cirad, Avenue Agropolis, 34398 Montpellier Cedex 5, France
(2) UMR 1097, Diversité et Génome des Plantes Cultivées, INRA, Domaine de Melgueil, 34130 Mauguio, France.
(3) UMR 5171, Génome Populations Interactions Adaptation, Université Montpellier 2, CC 63 Bât. 24, place Eugène Bataillon, 34095 Montpellier, cedex 5

OBJECTIFS DES RECHERCHES

Les objectifs du projet étaient, pour un ensemble de 5 espèces aux caractéristiques biologiques différentes, (*i*) de déterminer l'effet de la structure génétique des échantillons sur l'étendue du DL et (*ii*) d'évaluer la faisabilité d'études d'association au sein des collections de ressources génétiques établies pour ces espèces. Au-delà de ces questions générales, les différents cas étudiés fournissent un ensemble d'exemples méthodologiques pour le développement de marqueurs génétiques pour l'étude du déséquilibre de liaison, et adaptés à des situations variées (information plus ou moins précises sur le génome, régions génomiques simples- versus multi- copies, …).

I. PRESENTATION DES TRAVAUX

INTRODUCTION – PROBEMATIQUE GENERALES

Des progrès considérables ont été effectués chez les plantes cultivées en terme de localisation sur le génome des gènes d'intérêt (gène majeurs et QTLs) en utilisant des populations bi-alléliques développées dans cet objectif. Toutefois, le nombre limité d'évènements de recombinaison existant dans des populations de ce type conduit à une faible résolution de la position des QTLs, même avec de grands effectifs. Des centaines de gènes peuvent se rencontrer dans une zone de quelques centimorgans, correspondant à l'intervalle de confiance classiquement détecté dans les recherches de QTLs, ceci rendant l'identification du gène concerné difficile. De plus, chacune de ces études n'échantillonnant qu'un faible nombre d'allèles, la grande diversité génétique disponible existant au sein de ces espèces n'est pas prise en compte.

Les généticiens humains conduisent depuis de nombreuses années des études de déséquilibre de liaison (DL) dans des populations naturelles. Ils ont montré que le DL, pour peu qu'il ne s'étende pas sur une trop grande distance, pouvait être utilisé pour localiser beaucoup plus précisément les gènes de maladie (Kruglyak, 1999) par la simple comparaison d'individus sains et malades.

Afin de pouvoir localiser, grâce à des études d'association, des gènes d'intérêt, il est nécessaire d'avoir un maillage très fin du génome (un marqueur tous les 6 kb chez l'homme par exemple ce qui correspond à environ 500 000 SNPs) pour avoir des valeurs de DL utilisables quel que soit le point du génome considéré ou d'étudier le DL dans des zones préalablement identifiées (gènes candidats, par exemple). Cependant, en raison de la domestication (fondation, sélection humaine) exercée sur les plantes, on peut espérer une étendue du DL supérieure à celle observée en génétique humaine.

Peu d'études de DL ont été conduites à ce jour sur des espèces végétales. Elles concernent principalement Arabidopsis, espèce pour laquelle les travaux de séquençage et les données produites par la génomique sont considérables, et le maïs.

Les premiers résultats mesurant l'intensité du DL ont été obtenus chez la canne à sucre (Jannoo et al, 1999), la betterave (Kraft et al, 2000), le maïs (Remington et al, 2001; Tenaillon et al, 2001; Ching et al, 2002), le blé (Pestsova and Roder, 2002) et Arabidopsis (Nordborg et al, 2002; Hagenblad and Nordborg, 2002). Ces études portent sur une ou plusieurs régions du génome. Elles font apparaître une étendue de DL très variable, en fonction des espèces. Le DL est d'environ 10cM chez les cultivars modernes de canne à sucre (espèce allogame à multiplication végétative) issus de la domestication récente de cette espèce et d'un petit nombre de méioses, de 250 Kb chez Arabidopsis (espèce autogame non cultivée) et de 200 pb à 1500 pb chez le maïs (espèce allogame à domestication ancienne). Ces études ont aussi montré une grande variabilité de l'étendue du DL en fonction de la région du génome et de la population étudiée: un DL s'étendant sur plus de 100 kb a été observé dans une population de lignées élites de maïs (Ching et al, 2002). Une étude d'association caractères-marqueurs conduite chez le maïs, a permis d'identifier des SNPs situés dans le gène dwarf 8 (Thornsberry et al, 2001) montrant des associations significatives avec la date de floraison et montré l'intérêt de ce type d'approche chez les espèces végétales. La plupart de ces études de DL ont été conduites en étudiant le polymorphisme existant dans et autour de gènes candidats, au moyen de SNPs

On s'attend à ce que la structure du DL dépende (*i*) du système de reproduction de l'espèce étudiée (autogamie/allogame), mais aussi (*ii*) de la région génomique analysée (taux de recombinaison local, sélection) et enfin (*iii*) de l'histoire de l'échantillon considéré (population/collection). La masse de données actuellement disponibles chez l'homme indique en effet une grande hétérogénéité de la structure du DL même au sein de la même espèce (Pritchard et Przeworski, 2001). Chez les plantes, peu de données sont disponibles, notamment pour des espèces autogames (Flint-Garcia, 2003). C'est cependant chez ces espèces que le ratio entre niveaux de polymorphisme et taux de recombinaison efficace est attendu le plus favorable pour une étude fine du déséquilibre de liaison (Nordborg, 2000).

Dans ce projet, il était proposé d'étudier le déséquilibre de liaison (DL) présent au sein de collections de ressources génétiques de cinq espèces de plantes, annuelles ou pérennes, autogames ou allogames (*Medicago truncatula*, blé dur, riz, sorgho et hévéa). Au niveau technique, l'intérêt des marqueurs microsatellites devait être évalué, en terme de facilité d'obtention de marqueurs très liés, ainsi que de capacité à révéler le DL. Parallèlement, différents types de marqueurs (SNP, RFLP et microsatellites) devaient être comparés sur les

mêmes matériels. Ainsi, à partir de ces 5 modèles biologiques différents, ont se proposait d'aborder des questions méthodologiques d'analyse du DL sur trois niveaux :

1. Faisabilité du développement ciblé de marqueurs (microsatellites alias SSR, SNP)
2. Influence de la nature des marqueurs (diversité allélique, type de motif SSR…) sur la mesure du DL
3. Influence de la structure génétique des populations d'étude sur le DL

Références :

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics 3:19-25

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev. Plant Biol, 54:357-374

Jannoo N, Grivet L, Dookun A, D'Hont A, Glaszmann JC (1999) Linkage disequilibrium among modern sugarcane cultivars. Theor Appl Genet, 99:1053-1060

Hagenbald J, Nordborg M (2002) Sequence variation and haplotype structure surrounding the flowering time locus FRI in Arabidopsis thaliana. Genetics, 161:289-298

Kraft T, Hansen M, Nilsson NO (2000) Linkage disequilibrium and fingerprinting in sugar beet. Theor Appl Genet, 101:323-326

Kruglyak, J (1999) Prospect for whole genome linkage disequilibrium mapping of common disease genes. Nat Genet, 22:139-144

Nordborg M.(2000) Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization. Genetics, 154: 923–929.

Nordborg M, Borevitz JO, Begerlson J, Berry CC, Chory J (2002) The extent of linkage disequilibrium in Arabidopsis thaliana. Nature Genetics, 20:190-193

Pestsova E, Roder M (2002) Microsatellite analysis of wheat chromosome D allows the reconstruction of chromosomal inheritance in pedigree of breeding programmes. Theor Appl Genet, 106:84-91

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet, 67:170-181

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet, 69(1): 1-14.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479-11484

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize. Proc Natl Acad Sci USA 98:9161-9166

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet, 28:286-289

RESUME (*POUR L'ENSEMBLE DU PROJET****)***

L'analyse de la liaison statistique entre allèles à 2 locus, ou déséquilibre de liaison (DL), a été abordée simultanément chez le riz, le sorgho, le blé, l'hévéa et Medicago truncatula *en utilisant des échantillons tirés de collections de ressources génétiques. Des stratégies utilisant des BACs ont permis de développer des marqueurs dans des zones localisées. L'utilisation des séquences terminales des BACs ciblés a été efficace pour définir des marqueurs microsatellites. Par ailleurs, des amorces ont été développées pour séquencer directement des fragments géniques. Le haut niveau de polymorphisme des marqueurs microsatellites a pu nécessiter des adaptations pour le calcul du DL. La structure génétique observée sur les plantes cultivées génère un fort bruit de fond de DL, tandis que dans les collections d'espèces sauvages autogames, le DL reste faible même à courte distance, décroissant rapidement au delà de 20 kb. A l'intérieur des groupes définis pour les espèces cultivées, le DL est plus faible et décroît assez rapidement, parfois de manière différente selon les groupes, de 100 kb environ pour le riz à 500 kb pour le sorgho. Une bonne connaissance de la diversité de l'espèce apparaît donc nécessaire avant de réaliser des études d'associations qui paraissent dès lors réalisables en utilisant les accessions contenues dans les collections de ressources génétiques.*

PRESENTATION DES TRAVAUX PAR ESPECE / EQUIPE

Une synthèse générale des résultats de l'ensemble du projet a été présentée lors du colloque BRG 2006 (La Rochelle, 2 Octobre 2006) :

Seguin M., Attard A., Bataillon T., Billot C., Cenci A., Chantret N., Courtois B., David J., Deu M., El Azhari N., Glaszmann J.-C., Glemin S., Haudry A., Le Guen V., Maynadier M., Pomies V., Ronfort J., Tsitronne A., Weber C. Analyse et prédiction des patrons de déséquilibre de liaison dans les collections de ressources génétiques de plantes pérennes ou annuelles, autogames ou allogames. *In : Les Actes du 6ème Colloque National du BRG. Ressources Génétiques : des Ressources Partagées*, La Rochelle, 2-4 October 2006, BRG publ., Paris, France, pp. 57-74.

Le présent rapport donne des informations plus détaillés sur les travaux conduits sur les 5 espèces du projets.

HEVEA (*Hevea brasiliensis*)

Introduction :

L'hévéa (*Hevea brasiliensis*) est une espèce pérenne allogame, originaire d'Amazonie, mais cultivée principalement en Asie du Sud-Est pour la production de caoutchouc naturel (latex). La cartographie/QTL a été développée chez cette espèce pour le marquage de facteurs de résistance à une maladie [1, 2], mais les difficultés de mise en œuvre, rendent particulièrement intéressantes les perspectives offertes par l'analyse du DL et la génétique d'association.

Le but de la présente étude était en premier lieu de développer des marqueurs fortement liés pour l'analyse du DL dans des zones cibles du génome de l'hévéa. Le développement de marqueurs SNP n'étant pas envisageable à court ou moyen terme, le choix s'est porté sur le développement de marqueurs SSR. L'intérêt de ces marqueurs en terme de polymorphisme et de répartition sur l'ensemble du génome avait été démontré chez l'hévéa ([3], [4] et données non publiées). De plus, des études réalisées sur des espèces modèles [5, 6] montraient que la densité en séquences SSR dans le génome des plantes, notamment dans des régions simples copies, devaient être suffisante pour permettre, avec une bonne probabilité, l'identification de marqueurs SSR à partir de séquences partielles de clones BAC par exemple. Le développement de marqueurs SSR à partir de telles séquences avait été réalisé chez plusieurs espèces [5, 7-9], mais le plus souvent à partir de BAC sélectionnés au hasard et en utilisant des étapes de sous-clonage des BAC et après présélection de BAC contenant des SSR.

L'objectif du cette étude était de tester sur hévéa : 1) la probabilité d'identification de SSR par simple séquençage d'extrémités de clones BAC, sans étape de sous clonage ou de présélection ; 2) la différence de densité en SSR entre des BAC sélectionnés au hasard et des BAC correspondant à des régions simples copies du génome ; 3) la faisabilité de cette stratégie de développement de marqueurs SSR liés pour l'analyse du DL.

Matériel & Méthodes :

La stratégie de développement ciblé de marqueurs SSR est illustrée figure hevea-1. La banque BAC hévéa [10] a été criblée à l'aide 16 sondes génomiques PstI, d'1 sonde ADNc homologues et 5 sondes homologues de gènes de résistance (Rga), choisies d'après leur localisation sur la carte de référence du génome de l'hévéa [11] et donnant des profils RFLP de locus non dupliqués. Les inserts des clones BAC positifs ont été extraits suivant un protocole standard et séquencés à partir des 2 extrémités (sous-traitance, Genome Express, France). Les motifs SSR retenus sont des répétitions de 1 à 6 bases, parfaites ou imparfaites, d'au moins 12 bases de longueur [6]. Les amorces PCR ont été définies à l'aide du logiciel Primer 3. Les marqueurs SSR ont été testés et révélés par électrophorèse en gel de polyacrylamide à 6,5 % et marquage radioactif (^{33}P). Le polymorphisme des SSR issus de BAC a été testé sur les 4 accessions d'hévéa, utilisées comme géniteurs des 3 populations de cartographie disponibles et la localisation chromosomique des SSR polymorphes a été confirmée par cartographie sur ces descendances.

Résultats :

Les résultats d'identification de motifs SSR, dans les extrémités de 37 BAC sélectionnés au hasard et de 87 BAC identifiés positifs par hybridation avec les 22 sondes cartographiées, sont donnés tableau *hévéa-1*. Conformément à notre hypothèse initiale, la densité en SSR apparaît 4 fois plus élevées dans des BAC correspondants à la fraction simple copie du génome que sur l'ensemble du génome de l'hévéa. La densité pour les BAC sélectionnés au hasard (distance moyenne de 6,1 kb entre 2 SSR) est comparable à la densité de 6-7 kb observée par Cardle et al. [5] chez différentes espèces de plantes. Par contre, la densité environ 4 fois plus élevée dans les BAC sélectionnés par hybridation (1,6 kb) a effectivement permis d'identifier des SSR pour 21 des 22 sondes testées, malgré le petit nombre de clones BAC positifs par sonde et la faible longueur des séquences analysées (650 bases en moyenne).

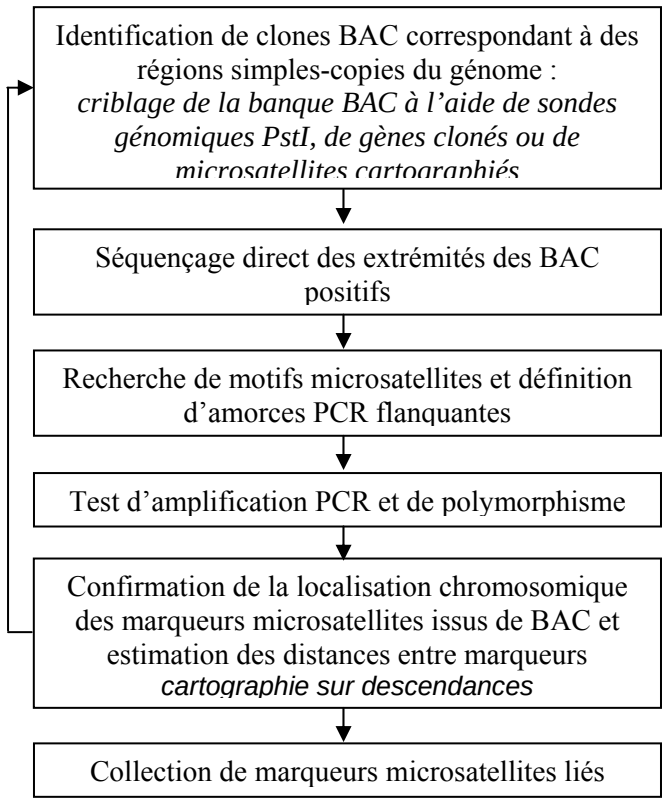


Figure hevea-1 : Stratégie d'identification ciblée de marqueurs microsatellites (SSR) à partir de ressources BAC.

Tableau hévéa-1 : Identification ciblée de marqueurs SSR chez l'hévéa : nombre de séquences analysées et densité en SSR ; comparaison entre des BAC sélectionnés au hasard et des BAC correspondant à des séquences non répétées.

	séquençage			**SSR identifiés***		
sélection des BAC	**clones BAC**	**extrémités de BAC**	**longueur séquencée (kb)**	**extrémités de BAC avec ≥ 1 SSR**	**SSR**	**distance moyenne (kb)**
hasard	37	55	30,5	3	5	6,1
hybridation°	87	141	91,6	38	56	1,6

* tous motifs mono-, di-, tri-, tetra- ou penta-nucléotidiques, parfait ou imparfaits, d'au moins 12 bases de longueur
° 22 sondes

Tableau hévéa-2 : Identification ciblée de séquences et marqueurs SSR chez l'hévéa : nombre de SSR identifiés dans les extrémités de BAC sélectionnés par hybridation à l'aide de 22 séquences non répétées.

type de répétition*	**mono-**	**di-**	**tri-**	**tetra-**	**penta-**	**total**
nombre	19	16	15	5	1	56
distance moyenne (kb)	5	6	6	18	92	1,6
marqueurs polymorphes (i.e. cartographiés)	9	8	4	-	1	22

* parfaites ou imparfaites, d'au moins 12 bases de longueur

Le tableau *hévéa-2* donne la répartition par longueur du motif répété des SSR identifiés et les résultats des tests de polymorphisme et de cartographie. Au final, des SSR polymorphes ont été obtenus pour 18 (13 PstI, 1 ADNc et 4 Rga) des 22 sondes sélectionnées confirmant la faisabilité de la stratégie proposée. Trois SSR polymorphes supplémentaires ne se sont pas cartographiés au même locus que la sonde correspondante (faux positifs) et n'ont pas été comptabilisés dans les 56 SSR identifiés au total. Au final, 22 marqueurs SSR ont été identifiés, *i.e.* des marqueurs qui se cartographient bien au locus attendu (co-localisation avec la sonde correspondante). La fréquence relative des différents motifs SSR dans les extrémités de BAC hévéa montre une proportion plus élevée de tri-nucléotidiques que ce qui est habituellement observé dans des régions intergéniques chez les végétaux [5, 12].

Discussion / perspectives :

Nos résultats démontrent la faisabilité à moindre coût du développement ciblé de marqueurs SSR. Le principe repose sur une forte densité de SSR dans la fraction simple copie – i.e. riche en gènes – du génome. La différence de taille du génome entre espèces végétales étant essentiellement due à la proportion de séquences répétées, cette méthode de développement ciblé devrait avoir une efficacité comparable quelque soit la taille du génome haploïde. L'application à l'hévéa, espèce au génome (2.1 pg/ 1C) environ 13 fois plus grand que celui d'*Arabidopsis*, semble confirmer cette hypothèse. La méthode proposée ici permet, de plus, l'identification de tous les motifs SSR présents dans le génome, alors que les méthodes par enrichissement se limitent à 1 ou quelques motifs. On pourra ainsi comparer le taux d'allélisme en fonction de la longueur du motif répété et l'incidence sur la mesure du DL. Les marqueurs SSR issus de BAC que nous avons identifiés seront utilisés en complément des SSR issus de banques enrichies et déjà cartographiés sur hévéa (280 SSR, données non publiées) offrant localement un jeu de marqueurs à des distances variables. Dans le cas où plusieurs SSR sont issu d'un même BAC ou d'un même contig, on dispose en plus d'une information sur la distance physique entre ces marqueurs.

Des difficultés techniques ont été rencontrées pour le séquençage des extrémités de BAC hévéa, ce qui a conduit à plusieurs mois de retard dans le développement des marqueurs. Ces problèmes ont été surmontés par d'autres équipes (Etienne Paux, comm. pers.) ce qui permet d'envisager l'application à haut débit de la méthode testée ici.

Le génotypage et l'analyse du DL à l'aide d'une cinquantaine de SSR sélectionnés d'après leur localisation et leur distance sur la carte génétique sont en cours sur une collection comprenant 350 accessions amazoniennes et 100 variétés cultivées.

Références :

1. Le Guen, V., et al., *Molecular mapping of genes conferring field resistance to South American Leaf Blight (Microcyclus ulei) in rubber tree.* Theoretical and Applied Genetics, 2003. **108**(1): p. 160-167.
2. Lespinasse, D., et al., *Identification of QTLs involved in the resistance to South American leaf blight (Microcyclus ulei) in the rubber tree.* Theoretical and Applied Genetics, 2000. **100**(6): p. 975-984.
3. Seguin, M., et al. *Microsatellite markers for genome analysis of rubber tree (Hevea spp.).* in *Proceedings of the IRRDB Symposium 2001 – Biotechnology & Rubber Tree*. 2002. 25-28 September 2001, Montpellier, France: Cirad, Montpellier, France.
4. Lekawipat, N., et al., *Genetic diversity analysis of wild germplasm and cultivated clones of Hevea brasiliensis Müell. Arg. using microsatellite markers.* Journal of Rubber Research, 2003. **6**(1): p. 36-47.
5. Cardle, L., et al., *Computational and experimental characterization of physically clustered simple sequence repeats in plants.* Genetics, 2000. **156**(2): p. 847-854.
6. Morgante, M., M. Hanafey, and W. Powell, *Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.* Nature Genetics, 2002. **30**(2): p. 194-200.
7. Allouis, S., et al., *Construction of a BAC library of pearl millet, Pennisetum glaucum.* Theoretical and Applied Genetics, 2001. **102**(8): p. 1200-1205.
8. Cregan, P.B., et al., *Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes.* Theoretical and Applied Genetics, 1999. **98**(6): p. 919-928.
9. Georgi, Y., et al., *Construction of a BAC library and its application to the identification of simple sequence repeats in peach [Prunus persica (L.) Batsch].* Theoretical and Applied Genetics, 2002. **105**(8): p. 1151-1158.
10. Piffanelli, P., et al. *BACTROP: a platform of genomic resources to study organization and evolution of tropical crop species*. in *7th International Congress of Plant Molecular Biology, June 23-28*. 2003. Barcelona, Spain: ISPMB.

11. Lespinasse, D., et al., *A saturated genetic linkage map of rubber tree (Hevea spp.) based on RFLP, AFLP, microsatellite, and isozyme markers.* Theoretical and Applied Genetics, 2000. **100**(1): p. 127-138.
12. Toth, G., Z. Gaspari, and J. Jurka, *Microsatellites in Different Eukaryotic Genomes: Survey and Analysis.* Genome Res., 2000. **10**(7): p. 967-981.

SORGHO (*Sorghum bicolor*)

Introduction

Le sorgho est une plante annuelle, essentiellement autogame, issue d'une domestication très ancienne en Afrique. On s'attend à ce que le DL soit restreint à des distances génétiques très faibles. Plusieurs dizaines de milliers de cultivars traditionnels sont présents dans les collections. Un gros travail de caractérisation morpho-agronomique et moléculaire a été réalisé, notamment au Cirad, et différentes formules de core collections ont été produites avec l'Icrisat.

Matériel et méthodes

Une core collection de 205 variétés représentative de la diversité de l'espèce cultivée *Sorghum bicolor* ssp. *bicolor* a été utilisée pour cette étude. Cette core collection, représentative de la diversité des races, latitudes, réponses à la durée du jour et systèmes de culture du sorgho a été analysée au moyen de 74 sondes RFLP réparties sur le génome [1]. L'analyse a montré une différenciation entre variétés africaines du Nord et du Sud de l'équateur (figure sorgho-2). La race est aussi un des facteurs impliqués dans la structuration de la diversité génétique des sorghos.

L'analyse du DL porte sur une région couvrant une distance génétique d'environ 5 cM, située à l'extrémité du bras court du chromosome 4 (figure sorgho-1). Cette région est synténique d'une région de la canne à sucre portant un gène majeur de résistance. Une carte physique partielle est disponible, constituée par deux contigs : un grand contig, constitué de 4 petits contigs, et d'une taille globale de 750 kb (y compris les zones de chevauchement) et un plus petit contig de 250 kb, ces 2 contigs, non chevauchants, étant séparés par une distance estimée de 200 kb. Douze marqueurs RFLP provenant de l'isolement des extrémités de BAC des contigs considérés ont été obtenus. La recherche et le développement de marqueurs microsatellites dans cette zone n'ont pas été couronnés de succès au cours de ce projet, que ce soit par hybridation d'oligonucléotides (CT) 15 et (GT) 15 sur les extrémités de 30 BACs choisis dans le grand contig de 750 kb et de 14 BACs dans le contig de 250 kb ou que ce soit par séquençage des extrémités. Aucun microsatellite nouveau n'a pu être obtenu dans cette zone. Toutefois, la recherche de microsatellites par construction de banques enrichies, spécifiques de BACs, avait permis le développement de 5 marqueurs microsatellites dans cette zone.

Les banques spécifiques enrichies en marqueurs microsatellites ont été produites à partir des 5 contigs de BAC identifiés (une banque regroupant des BACs couvrant l'ensemble de chaque contig) et non à partir de BAC isolés. Aussi, la position des microsatellites n'a pu qu'être estimée, seuls les marqueurs RFLP sont ordonnés avec sûreté sur les contigs car leur cartographie physique a pu être réalisée. Toutefois, pour le positionnement des microsatellites ont été pris en compte d'une part, les redondances entre séquences appartenant à des contigs différents (92 séquences obtenues) et d'autre part, la position des marqueurs RFLP sur le contig (si un marqueur RFLP est situé à l'extrémité d'un contig, le microsatellite ne pourra se situer au-delà de celui-ci).

Le DL sera donc analysé, au sein de cette zone, à l'aide de 12 marqueurs RFLP et 5 marqueurs microsatellites, situés dans une zone couvrant environ 750 kb, et 4 cM. (figure sorgho-1).

Le DL a aussi été étudié dans une autre zone du génome. La séquence du clone BAC sorgho (130 kb) contenant le gène waxy étant disponible, nous avons développé de nouveaux marqueurs microsatellites au sein de ce BAC. Sept couples d'amorces permettant l'amplification de marqueurs microsatellites ont été définis et les données de génotypage ont été acquises pour 5 de ces microsatellites.

La structuration de la core collection a été analysée avec la méthode Bayesienne mise en œuvre dans le logiciel Structure, v. 2.1. Nous avons utilisé le modèle avec admixture et fréquences corrélées, avec un nombre de populations (K) variant de 1 à 12, avec 10 réplications pour chaque valeur de K, une «burning» période de 50 000 et 500 000 itérations. La simulation montrant la plus haute probabilité postérieure des données a été choisie pour chaque valeur de K.

Le DL mutiallélique a été mesuré (D' et r^2), pour chaque paire de marqueurs, avec le logiciel Tassel, V. 1.9.0. La significativité du test a été évaluée par 1000 permutations.

Résultats et discussion

Comparaison des indices D'et r2

La courbe des indices multi-alléliques D'et r^2 calculés entre les 17 marqueurs liés en fonction de la distance physique montre une faible décroissance du r^2 en fonction de la distance (figure sorgho-2). Nous avons calculé le DL entre marqueurs non liés (60 locus RFLP répartis sur le génome) et obtenu une forte variation de l'indice de r^2, de 0 à 0.51. Un DL fort, lié à la structuration de la core collection, existe entre marqueurs non liés. La distribution des r^2 entre ces marqueurs non liés a permis de déterminer une valeur seuil de 0.18 pour le r^2 (95% des DL calculés ont une valeur inférieure à ce seuil sur l'ensemble du génome). Pour les locus liés, la courbe montre que certains marqueurs séparés par des distances d'environ 400-500 kb présentent un DL fort, cas de R et J distants de 492 Kb et dont le r^2 est de 0.26 ainsi que de R et L, distants de 435 kb, r^2 de 0.4.

Comparaison des DL obtenus avec différents types de marqueurs

a) DL entre marqueurs microsatellites

Au sein du BAC « Waxy » :

Les 5 microsatellites situés sur le BAC contenant le gène waxy ont présenté un nombre d'allèles variant de 3 à 13, dans la core collection. Ils sont séparés par des distances variant de 2 à 97 kb. Ces microsatellites n'ont montré aucun DL significatif (r^2 maximal observé de 0.017 pour 2 marqueurs distants de 46 kb). Les 2 microsatellites les plus proches, distants d'environ 2 kb, et présentant respectivement 3 et 13 allèles, ont montré un DL non significatif et très faible (r^2 de 0.008).

Sur le chromosome 4, dans la zone étudiée (750 kb, 4 cM) :

Les 5 microsatellites ont révélé un nombre d'allèles variant de 3 à 17. Le DL calculé entre toutes ces paires de marqueurs est très faible (figure sorgho-3), le maximum observé est un r^2 de 0.16 pour 2 microsatellites distants d'environ 200 kb.

b) DL entre marqueurs RFLP et entre RFLP et microsatellites

Le DL le plus fort est observé entre les marqueurs RFLP avec un r^2 maximum de 0.67 pour des locus distants de 17 kb, et 12 couples de RFLP montrant un DL $\geq$ 0.3 (figure sorgho-3). Toutefois, certains marqueurs microsatellites montrent un DL fort avec des marqueurs RFLP, en particulier les valeurs de DL les plus élevées sont trouvées entre le microsatellite C4 et 3 marqueurs RFLP (la distance entre les couples variant de 75 à 567 kb, et le r^2 variant de 0.22 à 0.27) ainsi qu'entre le microsatellite F1 et 4 RFLP (distance variant de 20 à 263 kb et r^2 variant de 0.18 à 0.26). Ces marqueurs microsatellites présentant du DL avec les marqueurs RFLP ont respectivement 12 et 13 allèles, le troisième microsatellite E2, présentant 17 allèles n'apparaît en DL avec aucun marqueur RFLP.

Les 2 autres microsatellites n'ont que 3 allèles et parmi eux, seul le microsatellite C5 est en DL avec un marqueur RFLP situé à 88 kb.

Influence de la structure sur le DL

La projection des haplotypes obtenus avec les locus liés (12 RFLP) montre que souvent un haplotype est caractéristique d'un groupe génétique identifié avec les locus non liés (résultats non montrés), ce qui indique une forte influence de la structuration génétique sur le DL entre locus liés. L'analyse sur les 60 locus non liés avec le logiciel Structure confirme la forte différenciation (Fst= 0,29) en 2 grands pôles géographiques identifiés par l'analyse Neighbor Joining, correspondant aux variétés africaines originaires du Nord de l'équateur et à celles originaires du Sud de l'équateur, les variétés asiatiques se répartissant dans les 2 grands pôles (figure sorgho-4). A la probabilité seuil de 70%, 129 variétés sont classées dans le groupe 1, Nord de l'équateur, (H = 0395, richesse allélique: 2,68) et 51 variétés dans le groupe 2 (H = 0,237 ; richesse allélique: 1,91), 25 variétés apparaissant hybrides entre les 2 groupes.

Le DL entre locus liés a été calculé à l'intérieur de chacun des 2 groupes afin de réduire l'effet de la structure sur le DL entre locus liés.

Au sein du groupe 1, un nombre plus réduit de marqueurs apparaît en DL, par comparaison avec l'analyse effectuée sur la core collection globale. En particulier, des DL existant entre marqueurs distants d'environ 50 et 200 kb ne sont plus conservés au sein de ce groupe. Mais, un DL élevé entre 2 marqueurs distants d'environ 400 kb a été observé (figure sorgho-5). Le maintien du DL au sein de ce groupe sur d'assez longues distances peut encore être liée à une structuration au sein de ce groupe dans lequel 5 ou 6 sous-groupes ont été respectivement identifiés par l'analyse bayesienne conduite avec Structure et l'analyse Neighbor Joining.

Au sein du groupe 2, seuls 9 DL sont significatifs au seuil de 5% et un fort DL est révélé (r^2 de 0,48) entre 2 locus RFLP distants d'environ 500 kb. Des valeurs de r^2 de 1 ont été observées entre des plusieurs locus distants de 100 kb maximum, indiquant une histoire similaire de mutation et de recombinaison pour ces locus à l'intérieur de ce groupe et/ou un fort effet de goulot d'étranglement lors de la domestication. Cette dernière hypothèse semble en accord avec la domestication supposée plus récente de ces sorghos Africains du Sud de l'Equateur à partir d'un pool génétique réduit, confirmé par une forte réduction de la richesse allélique comparée à la core collection.

Conclusions

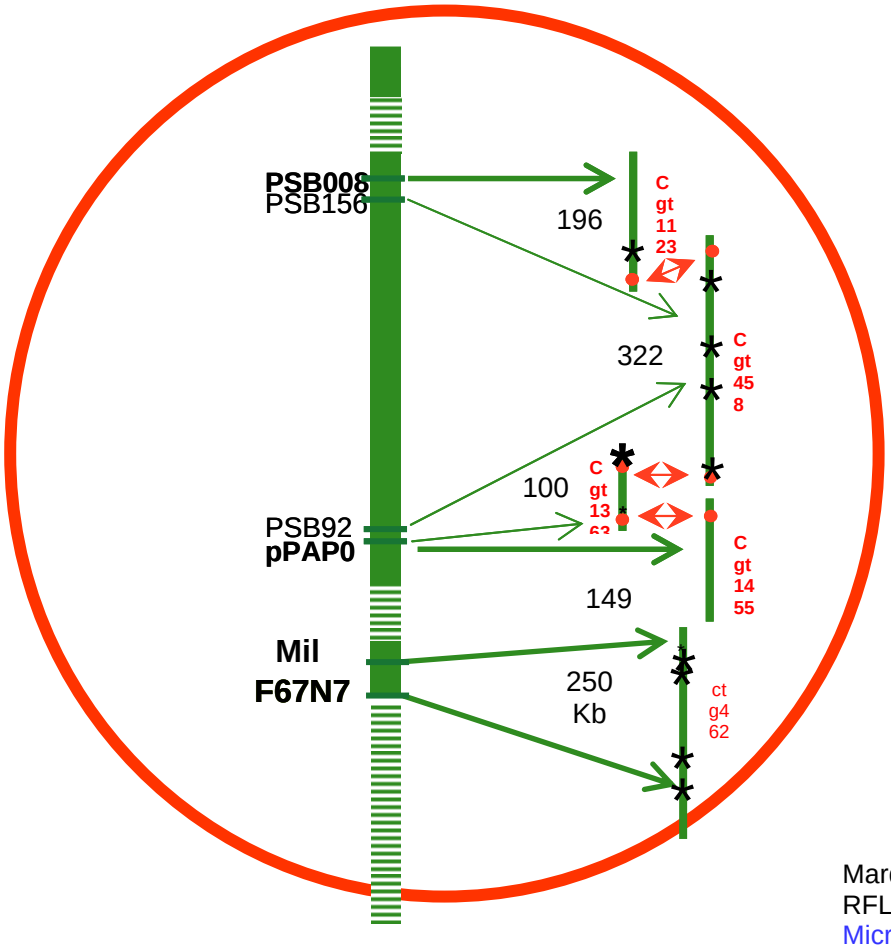
Certains marqueurs microsatellites ont montré leur capacité à révéler du DL avec des marqueurs RFLP, ce sont ceux possédant un nombre moyen d'allèles (environ 12). Par contre, aucun DL n'a été observé, au sein de la core collection, entre marqueurs microsatellites dans les 2 zones étudiées. Lorsque la diversité génétique est plus réduite dans une population (cas du groupe2), ils peuvent s'avérer des marqueurs pertinents pour conduire des études de DL. Les données de séquençage de plusieurs fragments du gène waxy, disponibles prochainement dans le cadre d'un autre projet, nous permettront d'étudier le comportement des microsatellites par rapport aux SNPs pour révéler du DL.

Un DL s'étendant jusqu'à 400-500Kb a été révélé bien qu'une forte variation des valeurs de DL existe avec, en particulier, des locus distants de 20 kb ne montrant aucun DL entre eux.

L'influence de la structure sur le DL a été testée mais nécessite un approfondissement. La race est aussi un facteur important dans la structuration de la diversité génétique des sorghos cultivés qui n'a pas encore été prise en compte dans les analyses de DL, en particulier au sein du groupe 1, comportant des groupes raciaux relativement différenciés.

Références :

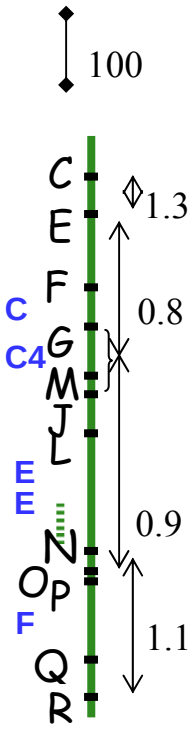
1. Deu M., Rattunde F., Chantereau J., A global view of genetic diversity in cultivated sorghums using a core collection, Genome 49 (2006) 168-180.
2. Pritchard J.K., Stephens M., Donnelly P., Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945-959.



Travaux de clonage positionnel dans la région homologue chez la canne à sucre

* Position des marqueurs RFLP dans chaque contig

Zone de chevauchement



Marqueurs utilisés pour l'étude du DL:
RFLP en noir
Microsatellites en bleu

Figure sorgho-1 : Carte des marqueurs utilisés pour l'analyse du DL chez le sorgho.

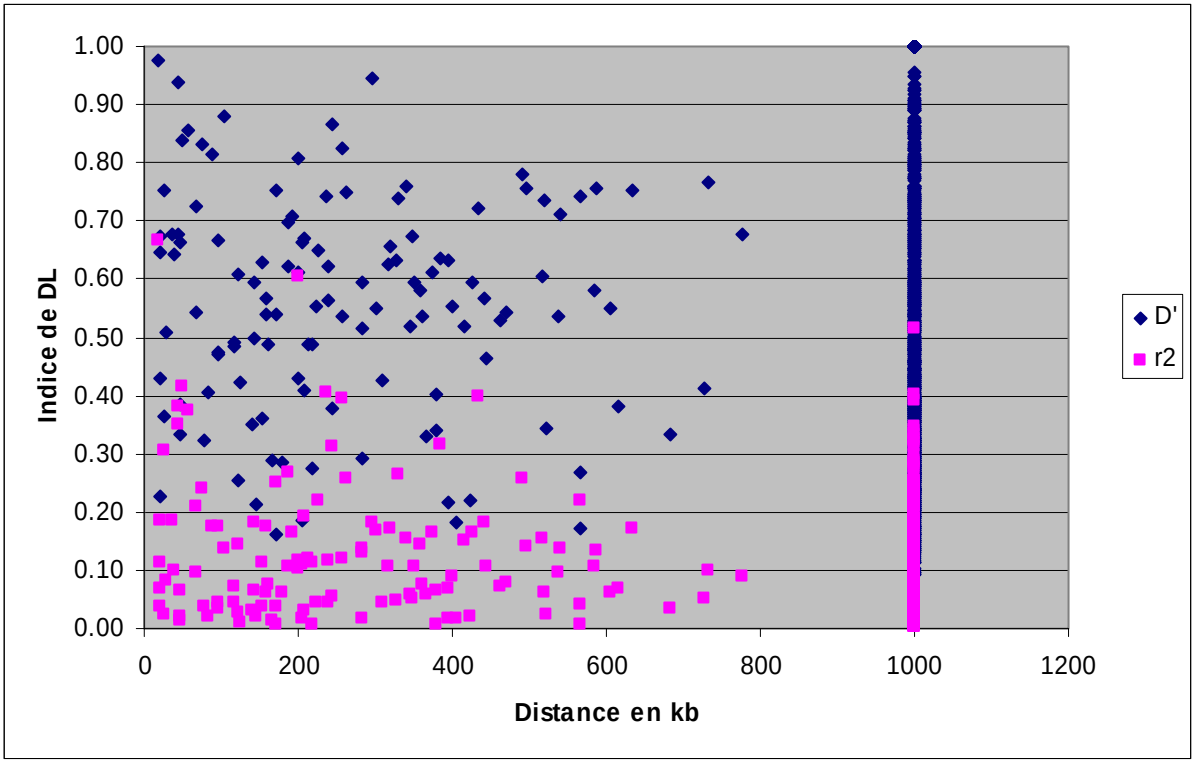


Figure sorgho-2: Evolution de D'et r^2 en fonction de la distance physique entre locus (situés sur le chromosome 4) au sein de la core collection (205 variétés); seules les valeurs significatives sont représentées ($p<0.05$). Les points à 1000 Kb correspondent aux valeurs de DL obtenus entre locus non liés (60 RFLP).

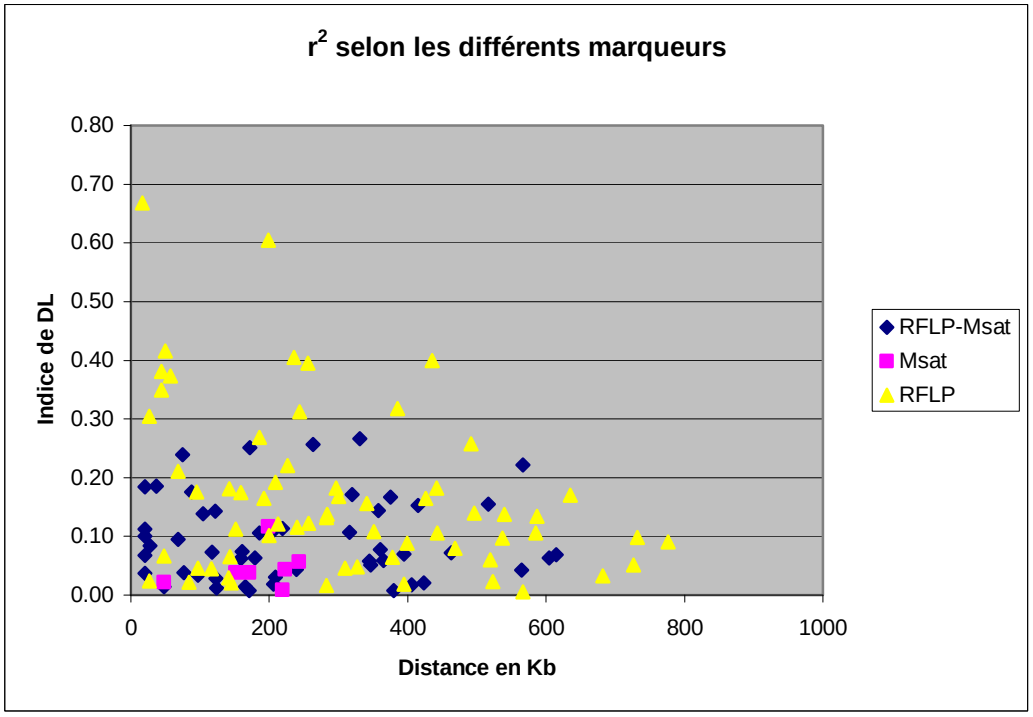


Figure sorgho-3 : Evolution de r^2 en fonction de la distance physique selon le type de marqueurs, au sein de la core collection, seules les valeurs significatives sont représentées ($p<0.05$).

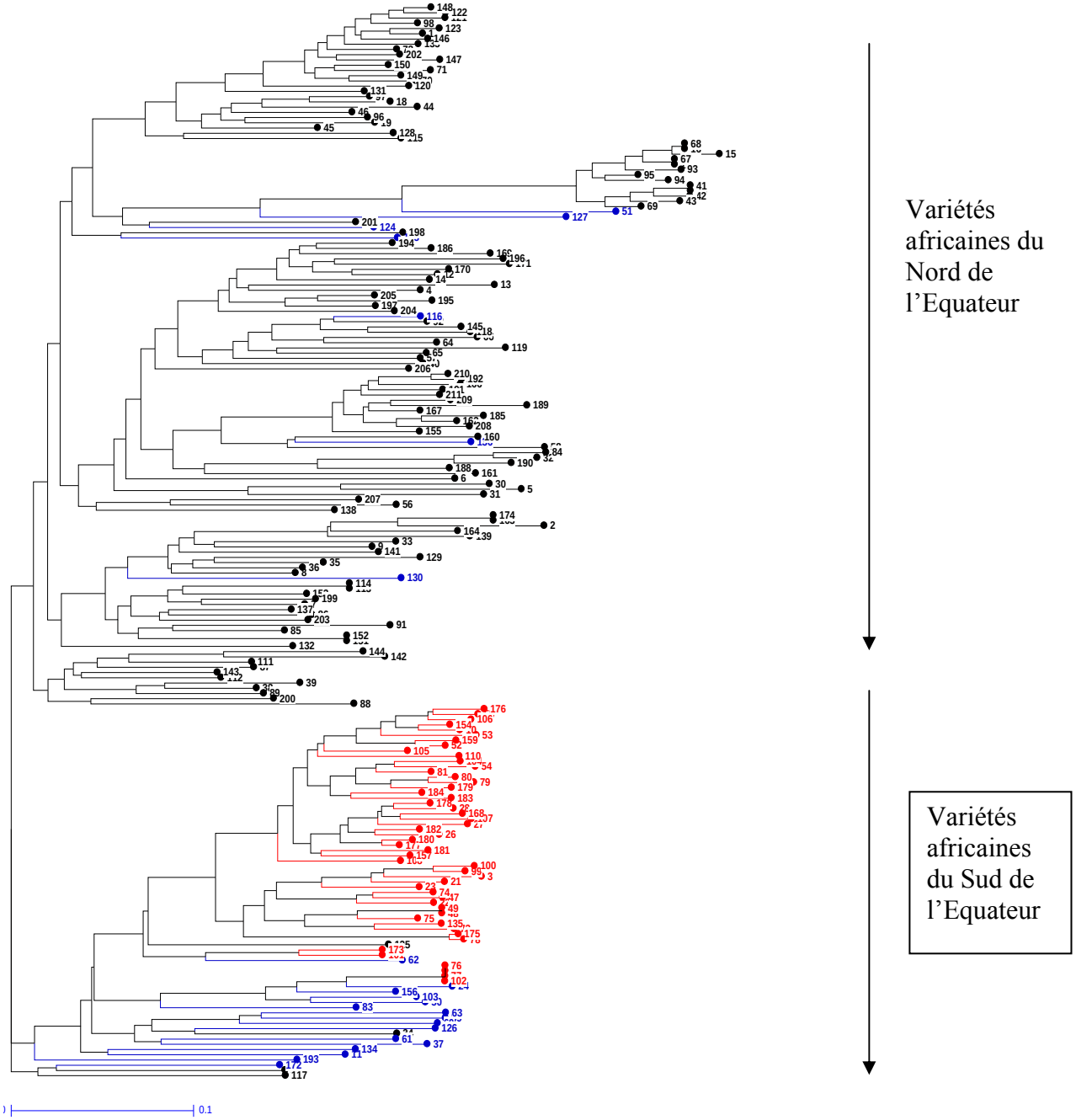


Figure sorgho-4: Structuration de la core collection obtenue avec les 74 marqueurs RFLP répartis sur le génome (NJ tree, indice de dissimilarité de Dice). En noir, sont représentées les accessions attribuées au groupe 1; en rouge les accessions attribuées au groupe 2; en bleu, les accessions «hybrides» entre les 2 groupes, une probabilité seuil de 70% ayant été choisi pour l'attribution à un groupe, et l'analyse conduite avec Structure [2] portant sur 60 locus non liés (distance minimale de 10 cM).

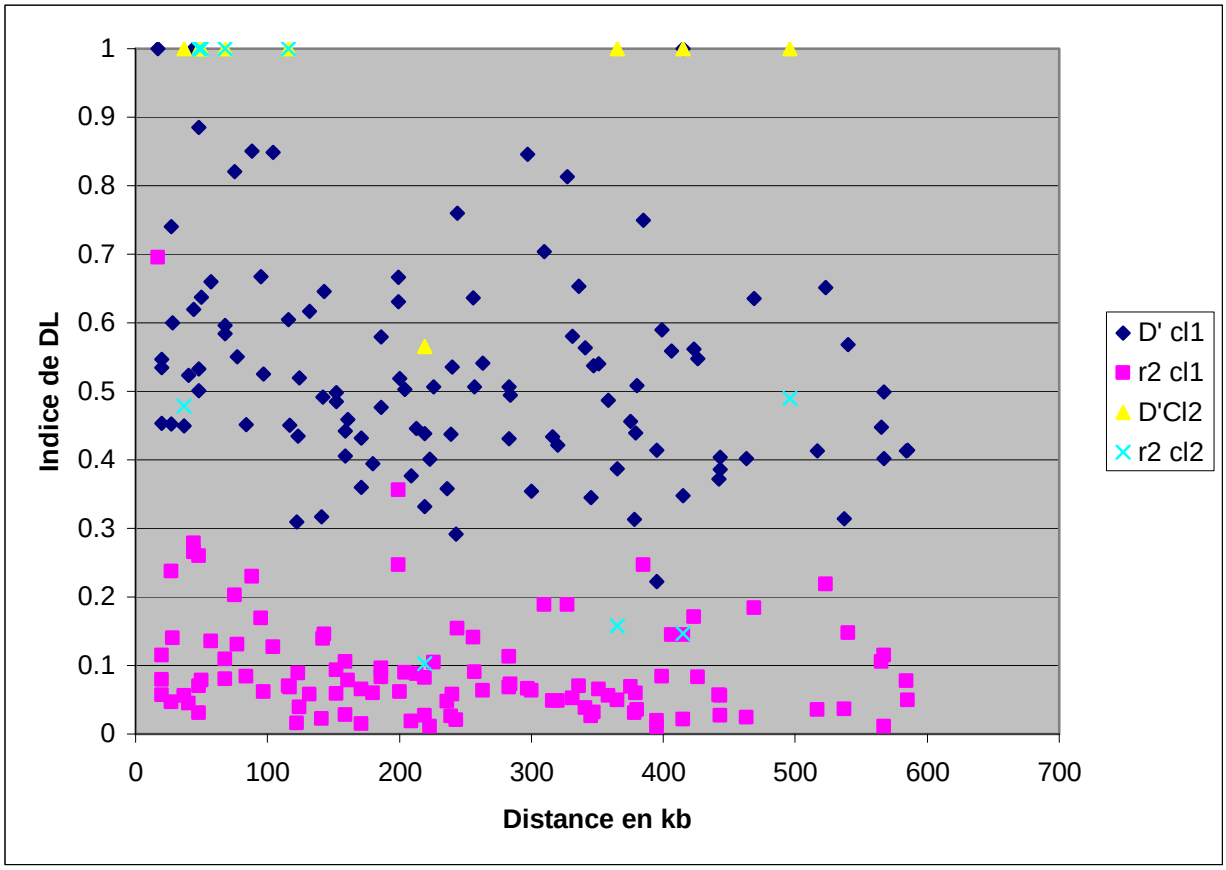


Figure sorgho-5 : Evolution de D'et r^2 en fonction de la distance physique entre locus (situés sur le chromosome 4), au sein des cluster 1 et 2 (définis par Structure), seules les valeurs significatives sont représentées ($p<0.05$).

MEDICAGO TRUNCATULA

INTRODUCTION

Medicago truncatula est une espèce autogame annuelle, dont l'aire de répartition naturelle est le bassin méditerranéen. D'importants efforts ont été réalisés pour rassembler et décrire la diversité génétique disponible dans les populations naturelles de cette espèce, aujourd'hui reconnue comme plante modèle pour la génétique et génomique des légumineuses. Une collection constituée de 400 lignées représentant l'ensemble de l'aire de répartition de *M. truncatula* a été constituée. Outre la description éco-géographique des sites de collecte, ce matériel a été caractérisé pour différents caractères phénotypiques et pour une vingtaine de marqueurs moléculaires et biochimiques (microsatellites et iso-enzymes). Ces données ont permis de structurer la collection et d'identifier différents groupes génétiques. Au titre de plante modèle, les régions riches en gènes de son génome font l'objet d'un séquençage systématique dans le cadre d'un consortium international. De nombreux BAC et contig de BAC sont ainsi disponibles (http://www.medicago.org/genome/) et permettent d'envisager une analyse de la relation « DL/distance physique » dans de nombreux points du génome. Dans le cadre de cette étude, l'analyse du polymorphisme de séquence et des patrons de déséquilibre de liaison (DL) devait nous permettre de (1) déterminer si la connaissance du DL au sein d'un échantillon permet de préciser ou de distinguer différents scénarios historiques [1] (par exemple repérer des situations d'admixture) et de (2) déterminer la faisabilité d'études d'association au sein d'une collection d'espèce autogame et l'effet de l'échantillonnage.

Pour appréhender l'effet de l'échantillonnage, nous avons fait porter l'analyse sur deux échantillons représentant différents niveaux d'apparentement : un échantillon représentant l'ensemble de la collection et potentiellement structuré et un échantillon issu d'un des groupes génétique détecter à l'aide des marqueurs microsatellites et supposé plus homogène. Pour estimer la vitesse de décroissance du DL en fonction de la distance physique, nous avons réalisé le séquençage de fragments balisant un contig de 120kb (figure Medicago-1).

Pour déterminer ce qu'apporte la connaissance du DL à l'étude de l'histoire démographique et génétique d'une espèce ou d'une population, des simulations basées sur la théorie de la coalescence ont été réalisées (logiciel ms [2]) et confrontées avec les patrons de polymorphisme et de DL observés dans les données.

Résultats-discussion :

Polymorphisme de séquence et déséquilibre de liaison

Cette étude donne un premier aperçu du polymorphisme de séquence au sein de l'espèce *M. truncatula* (voir cependant [3]). Comme observé chez *Arabidopsis thaliana*, le polymorphisme de séquence observé à l'échelle de l'espèce est relativement élevé ($\theta_S \sim 3.10^{-3}$). Les patrons de diversité observés pour les différents fragments étudiés ne montrent pas d'écart significatif au modèle neutre (équilibre mutation-dérive, [4]), malgré des valeurs du *D* de Tajima dans l'ensemble plutôt négatives. Ces valeurs négatives attestent d'un (faible) excès d'allèles rares à l'échelle de l'espèce, ce qui peut être interprété comme la signature d'une expansion démographique, de pressions de sélection positive ou d'un fonctionnement global de type métapopulation, associé à de forts taux de migration entre dèmes [5].

Malgré le fort taux d'autogamie de l'espèce, le DL observé est faible ($r^2<0.2$), décroît sur 10 à 20 kb et est peu dépendant du type d'échantillon considéré (figure Medicago-2). Les résultats sont néanmoins compatibles avec les attendus : le DL entre sites proches est plus fort dans l'échantillon structuré et la décroissance du DL plus rapide dans ce groupe. Le faible DL observé suggère que l'autogamie est une acquisition récente de l'espèce et/ou que la taille efficace de l'espèce est très élevée (des résultats similaires ont été obtenus chez *Arabidopsis thaliana*). Concernant la faisabilité des études d'association, nos résultats suggèrent qu'un maillage de l'ordre de 1 marqueur tous les 10 ou 20 kb pourrait permettre de réaliser des études d'association, et ceci pour les deux niveaux d'échantillonnage. Nos données sont cependant limitées à une très faible portion du génome et ne peuvent pas pour l'instant être généralisées.

Simulations :

Des algorithmes basés sur la théorie de la coalescence ont été réalisées pour construire des jeux de données (polymorphisme de séquence) attendus sous différents scénarios évolutifs (population en croissance,

goulot d'étranglement, population structurée) [2]. Ces jeux de données simulés ont été comparés aux données observées dans les différents échantillons pour essayer de déterminer le scénario expliquant le mieux nos données. .

Dans l'état actuel des analyses, il apparaît que les « statistiques » classiquement utilisées pour estimer le DL (r^2) ou pour faire des inférences sur le taux de recombinaison (R_{min}) n'apportent que peu d'information sur l'histoire démographique des populations. La comparaison du polymorphisme observé dans notre étude à celui attendu selon différents scénarios démographiques (population structurée, expansion démographique, goulot d'étranglement) suggère que les deux échantillons sont issus non pas d'une population subdivisée mais d'une population qui a connu une expansion démographique. Ces simulations ne prennent cependant pas en compte la dynamique d'extinction/recolonisation de l'espèce.

Références

1. Nordborg M., Tavare S., Linkage disequilibrium: what history has to tell us, Trends Genet. 18 (2002) 83-90.
2. Hudson R.R., Generating samples under a Wright-Fisher neutral model of genetic variation, Bioinformatics 18 (2002) 337-338.
3. de Mita S., Santoni S., Hochu I., Ronfort J., Bataillon T., Molecular evolution and positive selection of the symbiotic gene NORK in Medicago truncatula, J. Mol. Biol. 62 (2006) 234-244.
4. Tajima F., Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism, Genetics 123 (1989) 585-595.
5. Wakeley J., Aliacar N., Gene Genealogies in a Metapopulation, Genetics 159 (2001) 893-905.

6.

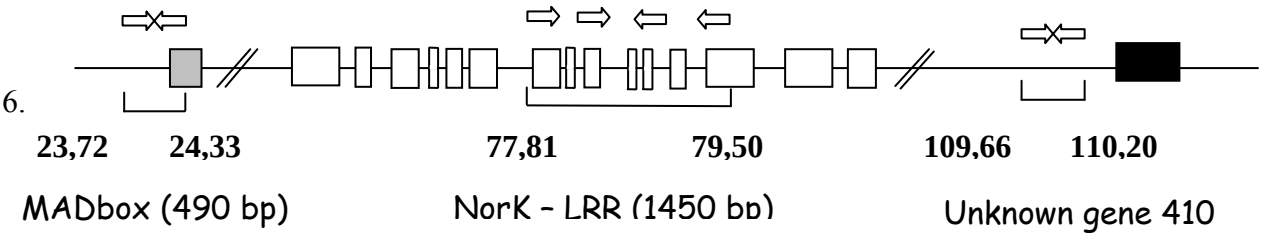


Figure Medicago-1 : Représentation schématique du BAC contenant le gène NorK (# AC126010) et des trois zones séquencées pour l'étude de la décroissance du déséquilibre de liaison. Les flèches signalent la position des amorces utilisées.

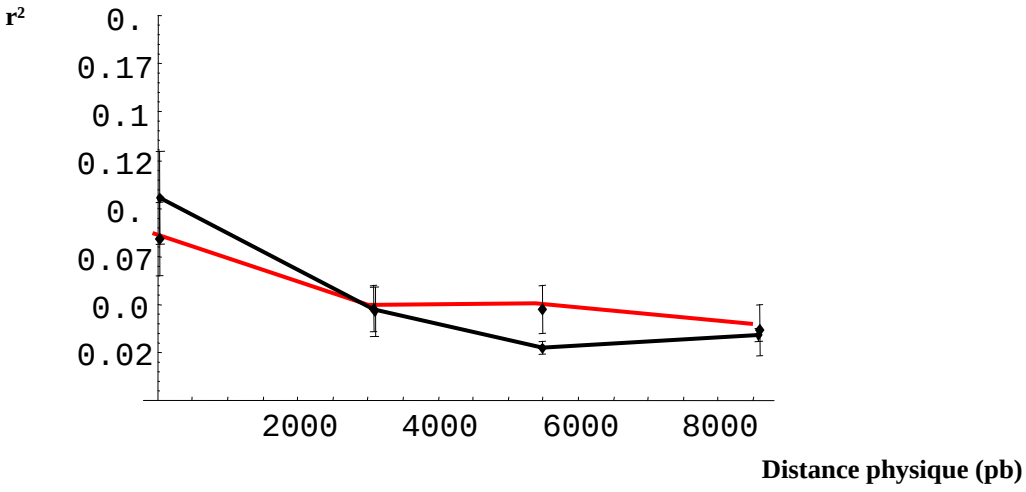


Figure Medicago-2 : Représentation graphique de la décroissance du déséquilibre de liaison mise en évidence autour du gène NorK. Le déséquilibre de liaison est estimé par r², la corrélation entre paires de sites polymorphes. La courbe noire représente les données du groupe global (potentiellement structuré); la courbe rouge les données de l'échantillon local (plus homogène). Les barres d'erreur donnent l'écart type (divisé par 10).

RIZ (*Sorghum bicolor***)**

-> RAPPORT DE STAGE D'INGENIEUR DE N. EL-AZHARI

BLE DUR (*Triticum turgidum*)

-> RAPPORT DE DEA D'A. HAUDRY

II. LISTE DES PRINCIPALES VALORISATIONS DES RECHERCHES

- Articles scientifiques

Deu M., Rattunde F., Chantereau J., A global view of genetic diversity in cultivated sorghums using a core collection, Genome 49 (2006) 168-180.

- Participations aux colloques nationaux ou internationaux (communication orale et poster).

Glaszmann J.C., Deu M. (2003). Analysing the genomic distribution of sequence diversity, the new challenge for germplasm characterization. 7th International Congress of Plant Molecular Biology, June 23-28. ISPMB, Barcelona, Spain, Abstract n° B-47: p22; présentée par B. Courtois.

Deu M. **,** Glaszmann J.C. (2004). Linkage disequilibrium in sorghum. Plant and Animal Genome VI Conference, January 10-14, 2004, San Diego, USA. présentée par JC Glaszmann

Deu M., Glaszmann J.C (2004). Selecting accessions for association studies on the basis of low-density whole-genome scans with molecular markers ; a case study with sorghum » . XVIIth EUCARPIA General Congress. Tulln, Autriche, 8-11 septembre 2004. présentée par M. Deu

- Rapports de fin d'étude (mémoires de maîtrise, de DEA, thèses...).

Annabelle Haudry (2004). Evolution du polymorphisme de séquence au cours de l'histoire des blés domestiques. DEA Ressources Phytogénétiques et Intéractions biologiques. AGRO-M, UM2, Montpellier, 40 p.

Najoi El Azhari (2004). Etude de l'étendue du déséquilibre de liaison local chez le riz (*Oryza sativa*). Cas des marqueurs microsatellites. Mémoire de fin d'études, Ingénieur des Techniques Agricoles de l'ENESAD, Dijon. 29 p.